Interpretation of likelihood

The likelihood function $L(\boldsymbol{\theta}; \mathbf{x})$ of a vector of parameters $\boldsymbol{\theta}$, based on a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, is a function of $\boldsymbol{\theta}$.

The meaning of $L(\boldsymbol{\theta}; \mathbf{x})$ is that its value gives a measure of "how likely" it is that $\boldsymbol{\theta}$ gives the true values of the parameter of interest, given the random sample $\mathbf{x}$.

Which $\boldsymbol{\theta}$?

If we take two different values of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, then a question arises on which of the two we should choose, given the sample $\mathbf{x}$.

The previous ideas might suggest that if

$$L(\boldsymbol{\theta}_1; \mathbf{x}) \geq L(\boldsymbol{\theta}_2; \mathbf{x}),$$

then $\boldsymbol{\theta}_1$ should be preferred, because $\boldsymbol{\theta}_1$ is "more likely" to describe the process underlying our experiment.

Maximum likelihood estimation

This idea leads to the principle of **maximum likelihood estimation**:

We estimate $\boldsymbol{\theta}$ by the value $\widehat{\boldsymbol{\theta}}$ which maximises the likelihood function, i.e.

$$\widehat{\boldsymbol{\theta}} \text{ is such that } L(\widehat{\boldsymbol{\theta}}; \mathbf{x}) \geq L(\boldsymbol{\theta}; \mathbf{x}), \quad \text{for all} \quad \boldsymbol{\theta} \in \Theta,$$

where Θ is set of possible parameters $\boldsymbol{\theta}$.

The maximum likelihood estimate

We call $\widehat{\boldsymbol{\theta}}$ the **maximum likelihood estimate** of $\boldsymbol{\theta}$, given the data $\mathbf{x}$.

Note that if we had different data $\mathbf{x}$, we would typically get a different $\widehat{\boldsymbol{\theta}}$. We hope that if we take enough samples, $\widehat{\boldsymbol{\theta}}$ becomes close to its true value $\boldsymbol{\theta}$.

Numbers of parameters

In some cases $\boldsymbol{\theta}$ will consist of just one parameter, in which case we say we have a **one-parameter problem**.

In some cases $\boldsymbol{\theta}$ will consist of two or more parameters, in which case we say we have a **multi-parameter problem**.

In the former case we can write $\boldsymbol{\theta} = \theta$ (a scalar parameter) and we will want to maximise $L(\theta; \mathbf{x})$ over θ.

In the latter case we will want to maximise $L(\boldsymbol{\theta}; \mathbf{x})$ over $\boldsymbol{\theta}$, a multi-dimensional maximisation problem.

Examples

Example 30: Discrete maximisation of likelihood

Example 31: Exponential maximum likelihood

Example 32: Binomial maximum likelihood

Maximising the likelihood

Maximum likelihood estimation comes down to a maximisation problem.

Whether this is easy or difficult depends on (a) the statistical model we use in the form $f(\mathbf{x}; \boldsymbol{\theta})$ and (b) the parameter vector $\boldsymbol{\theta}$.

One-parameter problems are clearly easier to handle and in many cases multi-parameter problems require the use of numerical maximisation techniques.

Log likelihood

In maximising $L(\boldsymbol{\theta}; \mathbf{x})$ it is usually easier to work with the logarithm of the likelihood instead of the likelihood itself.

We call the logarithm of the likelihood the **log-likelihood function** and we write

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \log L(\boldsymbol{\theta}; \mathbf{x}).$$

Maximising $\ell(\boldsymbol{\theta}; \mathbf{x})$ over $\boldsymbol{\theta}$ produces the same estimator $\widehat{\theta}$ as maximising the likelihood, because the logarithm is increasing, i.e.

$$\widehat{\boldsymbol{\theta}} \text{ is such that } \ell(\widehat{\boldsymbol{\theta}}; \mathbf{x}) \geq \ell(\boldsymbol{\theta}; \mathbf{x}), \text{ for all } \boldsymbol{\theta} \in \Theta.$$

In this course we work with natural logarithms, which are useful because many p.d.f.s include an exponential term.

The parameter set and maximisation techniques

When we maximise $\ell(\boldsymbol{\theta}; \mathbf{x})$, we need to be careful with the parameter set Θ.

In most of the examples we will meet in this module $\boldsymbol{\theta}$ will be continuous (NB this is not the same thing as saying that the distribution of $\mathbf{X}$ is continuous) and so we can use differentiation to obtain the maximum.

However, in some cases (e.g. Example 30) the possible values of $\boldsymbol{\theta}$ may be discrete (i.e. Θ is a discrete set) and in such cases we cannot use differentiation.

One parameter problems

One-parameter problems can be easily handled using the maximisation and minimisation techniques from single variable calculus theory.

For example to obtain the maximum of $\ell(\theta; \mathbf{x})$, we first find the solution of

$$\frac{d\ell(\theta; \mathbf{x})}{d\theta} = 0 \Rightarrow \theta = \widehat{\theta} \tag{1}$$

and then we check that

$$\left.\frac{d^2\ell(\theta; \mathbf{x})}{d\theta^2}\right|_{\theta=\widehat{\theta}} < 0. \tag{2}$$

Checking for maxima

Note that (1) only does not guarantee that $\widehat{\theta}$ is a maximum; it is necessary to check with (2).

(In some cases, the maximum likelihood estimate of θ does not exist!)

Examples

Example 33: Chemical reaction again

Example 34: Poisson maximum likelihood estimation

Example 35: Uniform maximum likelihood estimation

Multi-parameter problems

For multi-parameter problems, where $\boldsymbol{\theta}$ is a vector, a similar procedure can be followed.

Here for simplicity we consider only the case where there are 2 parameters (so that $\boldsymbol{\theta}$ is a 2×1 vector) and write $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$.

Stationary points

Now we find a stationary point $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2)^T$ of the log-likelihood by solving

$$\frac{\partial \ell(\boldsymbol{\theta}, \mathbf{x})}{\partial \theta_1} = 0, \quad \frac{\partial \ell(\boldsymbol{\theta}, \mathbf{x})}{\partial \theta_2} = 0. \tag{3}$$

Equation (3) is the analogue in the two parameter case of equation (1) in the one parameter case.

The Hessian

The candidate $\widehat{\boldsymbol{\theta}}$ may be a maximum or not, and we have to check this by using an analogue of equation (2) in order to check if $\widehat{\boldsymbol{\theta}}$ is indeed a (local) maximum of the log likelihood function.

First we calculate the so called **Hessian matrix**:

$$H = \begin{pmatrix} \partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})/\partial\theta_1^2 & \partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})/\partial\theta_1\partial\theta_2 \\ \partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})/\partial\theta_1\partial\theta_2 & \partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})/\partial\theta_2^2 \end{pmatrix}$$

and then we evaluate H at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, where $\widehat{\boldsymbol{\theta}}$ is the stationary point we found using (3).

Identifying maxima

If H is a negative definite matrix (the analogue of the second derivative being negative in the one parameter case), then $\widehat{\boldsymbol{\theta}}$ maximises $\ell(\boldsymbol{\theta}; \mathbf{x})$.

If H is not a negative definite matrix, then we cannot conclude that $\widehat{\boldsymbol{\theta}}$ is a (local) maximum.

Negative definite matrices

To check that H is a negative definite matrix we can use the following (in the 2 variable case): if

$$\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x}) / \partial \theta_1^2 < 0$$

and the determinant $\det(H)$ is positive, then H is negative definite.

(More detail on maximising and minimising functions of more than one variable can be found in the module MAS211 Advanced Calculus and Linear Algebra.)

Example

Example 36: Maximum likelihood estimation for normal distribution with unknown mean and variance